

## SUPPLEMENTARY MATERIAL

### Exome Capture from Saliva Produces High Quality Genomic and Metagenomic Data

Jeffrey M. Kidd<sup>1,2,\*</sup>, Thomas J. Sharpton<sup>3,4\*</sup>, Dean Bobo<sup>5</sup>, Paul J. Norman<sup>6</sup>, Alicia R. Martin<sup>1</sup>, Meredith L. Carpenter<sup>1</sup>, Martin Sikora<sup>1</sup>, Christopher R. Gignoux<sup>7</sup>, Neda Nemat-Gorgani<sup>6</sup>, Alexandra Adams<sup>1</sup>, Moraima Guadalupe<sup>8</sup>, Xiaosen Guo<sup>9</sup>, Qiang Feng<sup>9</sup>, Yingrui Li<sup>9</sup>, Xiao Liu<sup>9</sup>, Peter Parham<sup>6</sup>, Eileen G. Hoal<sup>10</sup>, Marcus W. Feldman<sup>11</sup>, Katherine S. Pollard<sup>3,12</sup>, Jeffrey D. Wall<sup>12</sup>, Carlos D. Bustamante<sup>1</sup>, Brenna M. Henn<sup>1,5,§</sup>

#### Supplementary Figures:

Figure S1 Pedigree structure for sequenced individuals  
Figure S2 Cumulative coverage across the Agilent target regions  
Figure S3 Mapping quality for all reads  
Figure S4 Assessment of base substitutions from mapped reads  
Figure S5 Venn Diagram comparing ≠Khomani San with Namibian exome samples  
Figure S6 PCA with two relatives included  
Figure S7 Distribution of mapped reads along the *N. subflava* genome  
Figure S8 The phylogenetic distribution of three non-human exome capture sequences that map with high fidelity to *Mycobacterium tuberculosis*.  
Figure S9 The phylum-level structure of the oral microbiome structure varies among the Khoesan.

#### Supplementary Tables:

Table S1 Summary Statistics for HGDP San Exomes  
Table S2 HLA data  
Table S3 KIR data  
Table S4 List of Taxa Used in Metagenomic Analysis (*separate file*)

## SUPPLEMENTARY MATERIAL

**Table S1** Summary Statistics for HGDP San Exomes

<b>Sample</b>	<b>Total Reads</b>	<b>Unmapped Reads</b>	<b>% Unmapped reads</b>	<b>% PCR Duplicates</b>	<b>Properly Paired</b>	<b>% Properly Paired</b>
HGDP00987	84,355,854	1,452,143	1.72%	9.66%	81,943,050	97.14%
HGDP00991	88,790,056	2,275,297	2.56%	7.85%	84,862,816	95.58%
HGDP00992	79,270,474	1,400,201	1.77%	9.85%	76,845,818	96.94%
HGDP01029	86,406,698	1,880,970	2.18%	9.78%	83,240,246	96.34%
HGDP01032	86,687,656	1,817,084	2.10%	10.28%	83,626,708	96.47%
HGDP01036	92,438,170	1,794,121	1.94%	11.31%	89,373,204	96.68%

## SUPPLEMENTARY MATERIAL

**Table S2:** *HLA alleles*

	HLA-A		HLA-B		HLA-C	
<b>SA006</b>	29:01	03:01	13:03	58:02	12:03	06:02
<b>SA008</b>	32:01	03:01	35:01	07:02	04:01	07:02
<b>SA011</b>	29:01	03:01	13:03	07:02	12:03	07:02
<b>SA012</b>	29:01	03:01	13:03	07:02	12:03	07:02
<b>SA035</b>	36:01	01:23	15:03	15:10	04:01	04:01
<b>SA051</b>	02:02	30:04	07:05	58:02	07:02	06:02
<b>SA052</b>	02:02	30:04	07:05	58:02	07:02	06:02
<b>SA054</b>	31:01	30:04	42:01	58:02	17:01	06:02
<b>SA1000</b>	33:01	68:02	41:01	53:01	04:01	17:01
<b>SA1001</b>	03:01	30:04	07:02	35:01	04:01	07:02
<b>SA1002</b>	11:01	30:01	35:01	42:01	04:01	17:01
<b>SA1006</b>	02:01	02:05	07:05	55:01	03:03	07:04
<b>SA1010</b>	03:01	68:01	58:01	58:02	06:02	x
<b>SA1025</b>	23:01	68:02	15:03	42:01	17:01	18:01

Shown are the alleles of HLA present in each of the individuals as determined from the exome sequences and verified by further sequencing (HLA-A, B, C). Alleles are named according to the ImmunoPolymorphism database (Robinson et al. 2010).

# SUPPLEMENTARY MATERIAL

**Table S3: *KIR* alleles**

	3DL3		2DS2		2DL2		2DL5B		2DS3/5c	
SA006	*038	*01405	*004	*001	2*003	2*001	*002	*002	3*007	3*001
SA008	*040	*01403		*001	3*005	2*003				
SA011	*01403	*038	*001	*004	2*003	2*003		*002		5*012
SA012	*01403	*01405	*001		2*003	2*001		*002		3*001
SA035	*037	*038	*001	*004	2*00601	2*009	B*018	B*00803	5*0502	5*012
SA051	*057	*038	*001		2*003	3*001	B*018		5*0502	
SA052	*024	*038	*001	*001	2*003	2*012		B*00803		5*012
SA054	*041	*020	*001		2*003	3*005	B*018		5*0502	
SA1000	*00301	*008	*001		2*003	3*001				
SA1001	*01303	*01403	*001		2*001	3*002	B*00803		S3*00103	
SA1002	*01403	*01405	*001	*001	2*003	2*003	B*002	B*00803	S3*00103	S5*006
SA1006	*00102	*038	*001		2*003	3*002	B*00801			S5*002
SA1010	*01303	*038	*001		2*003	3*002	B*002		S3*00103	
SA1011	*005	*037			3*001	3*018	B*00801			S5*002
SA1025	*00201	*038			3*001	3*001				

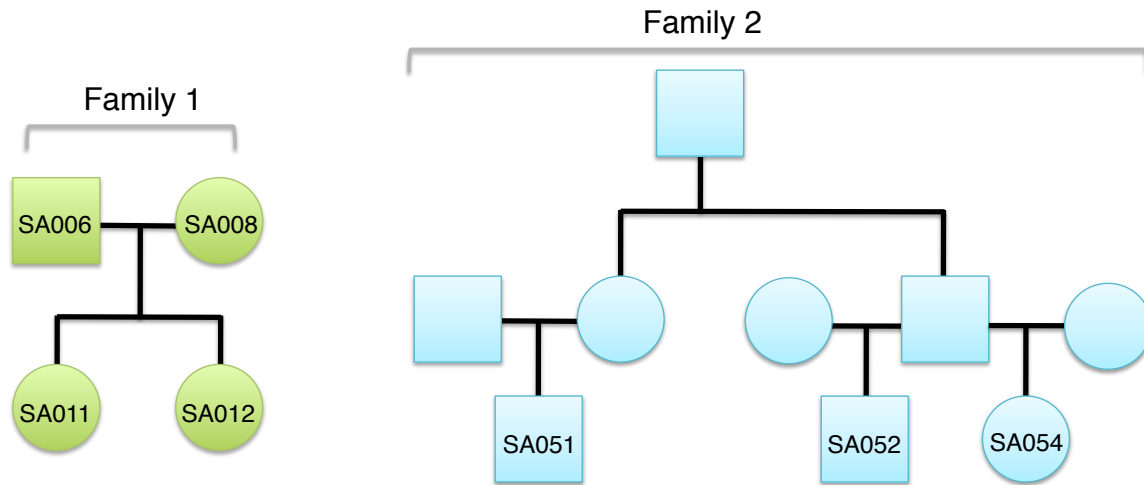
	2DL1		2DL4		3DL1/S1		2DS4		3DL2	
SA006	*00401	*00401	*00801	*006	*005	*007	*010	*004	*010	*008
SA008	*01202		*00103	*00802	*01502		*006		*00101	*054
SA011		*00401	*00802	*00801		*005		*010	*054	*010
SA012		*00401	*00802	*006		*007		*004	*054	*008
Sa035	*022		*013	*00103	*041	*070	*00101	*00101	*052	*00101
SA051	*022	*00302	*013	*00103	*041	*01502	*00101	*00101	*015	*063
SA052	*026N		*00102	*024	*002	*035	*00101	*00101	*00101	*015
SA054	*022	*00401	*027	*024	*041	*035	*00101	*00101	*052	*015
SA1000	*001	*004	*00102	*00103	*002	*070	*00101	*00101	*002	*023
SA1001	*001	*022	*00801	*013	*005	*041	*00101	*00101	*002	*015
SA1002	*010	*01201	*00801	*006	*007	*060	*00101		*008	
SA1006	*001	*01201	*00103	*00103	*01501	*008	*00101	*00101	*009	*053
SA1010	*002	*004	*00801	*01201	*001	*071	*00101	*00101	*001	*002
SA1011	2DS1*002		*00103		*020		*00101		*001	*006
SA1025	*001	*01201	*00801	*00801	*005	*060	*010		*010	

Shown are the KIR alleles present in each of the individuals as determined from the exome-capture sequences and verified by further sequencing. Alleles are named according to the ImmunoPolymorphism database (Robinson et al. 2010), and sequences discovered by the current study (shaded in purple) are denoted by local nomenclature. Clear boxes indicate that the gene is absent.



## SUPPLEMENTARY MATERIAL

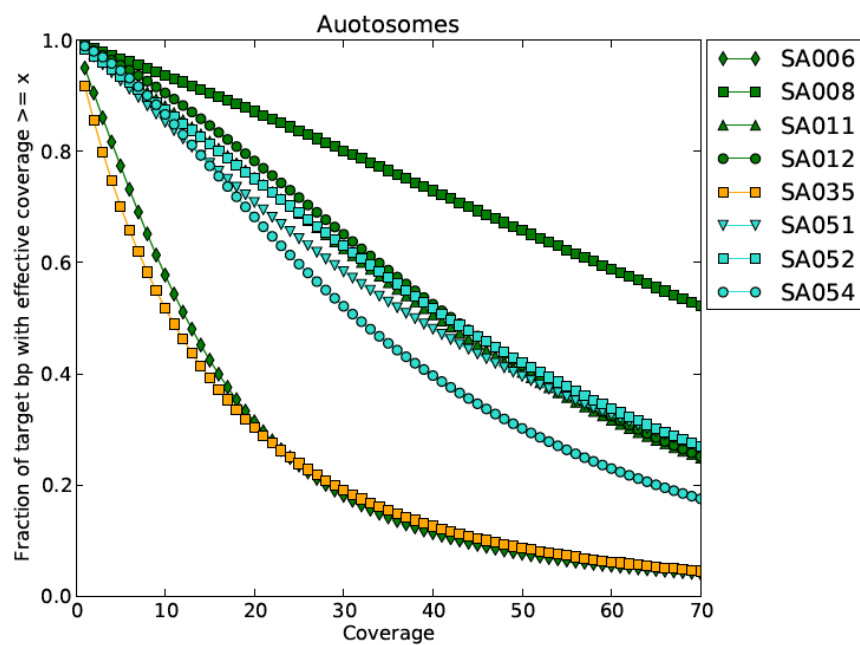
### Supplemental Figures:



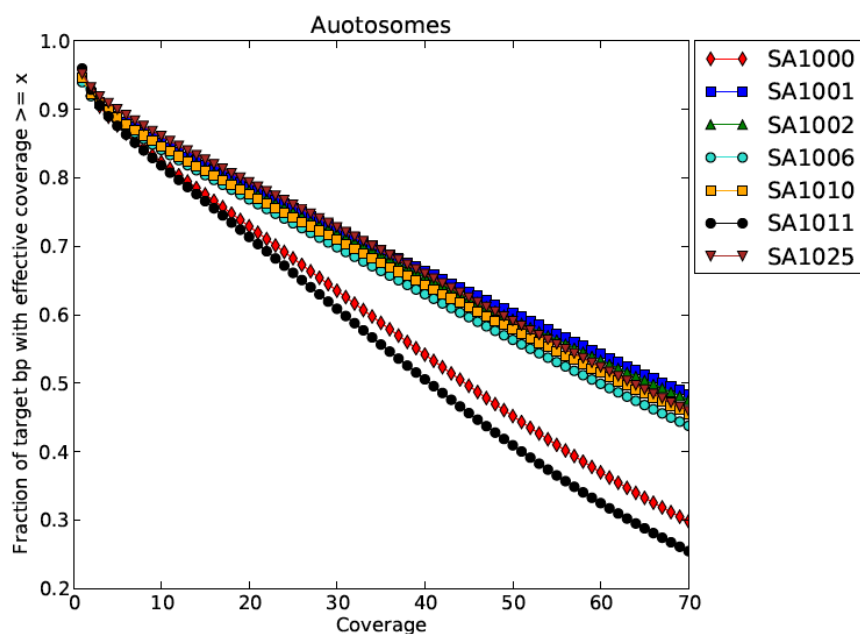
**Figure S1** *Pedigree structure for sequenced individuals.* We selected 15 KhoeSan saliva samples from the Northern Cape Province, South Africa for exome capture using the Agilent 50Mb and 44Mb exome kits. The sample set consisted of four individuals from family 1, a quartet with two daughters, three individuals from family 2, an extended family of half-siblings and first cousins, and 8 purportedly unrelated individuals. Individuals ranged in age between 30-100 years.

# SUPPLEMENTARY MATERIAL

A



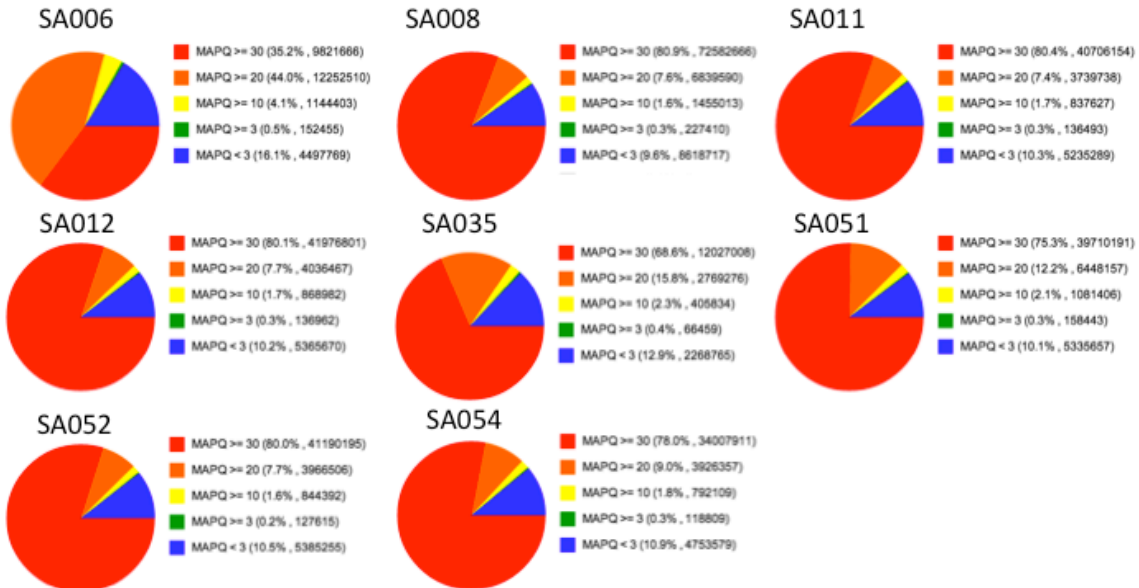
B



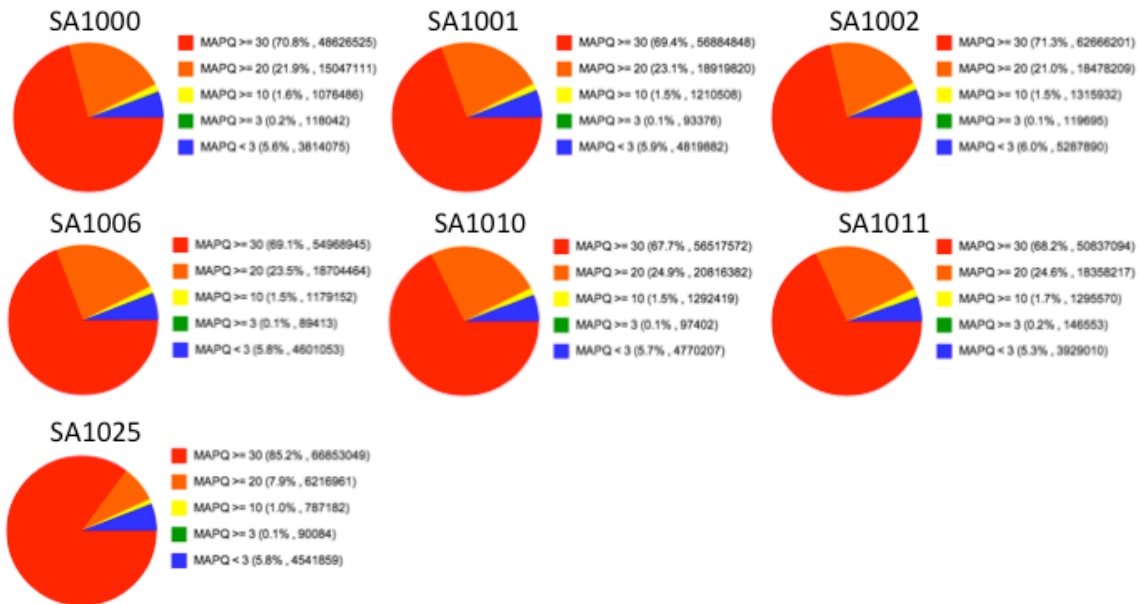
**Figure S2** Cumulative coverage across the Agilent target regions for Pilot 1 (A) and Pilot 2 (B) samples. Effective coverage is based on non-duplicate reads used for SNP discovery by the GATK Unified Genotyper. Analysis was limited to the autosomes.

## SUPPLEMENTARY MATERIAL

A

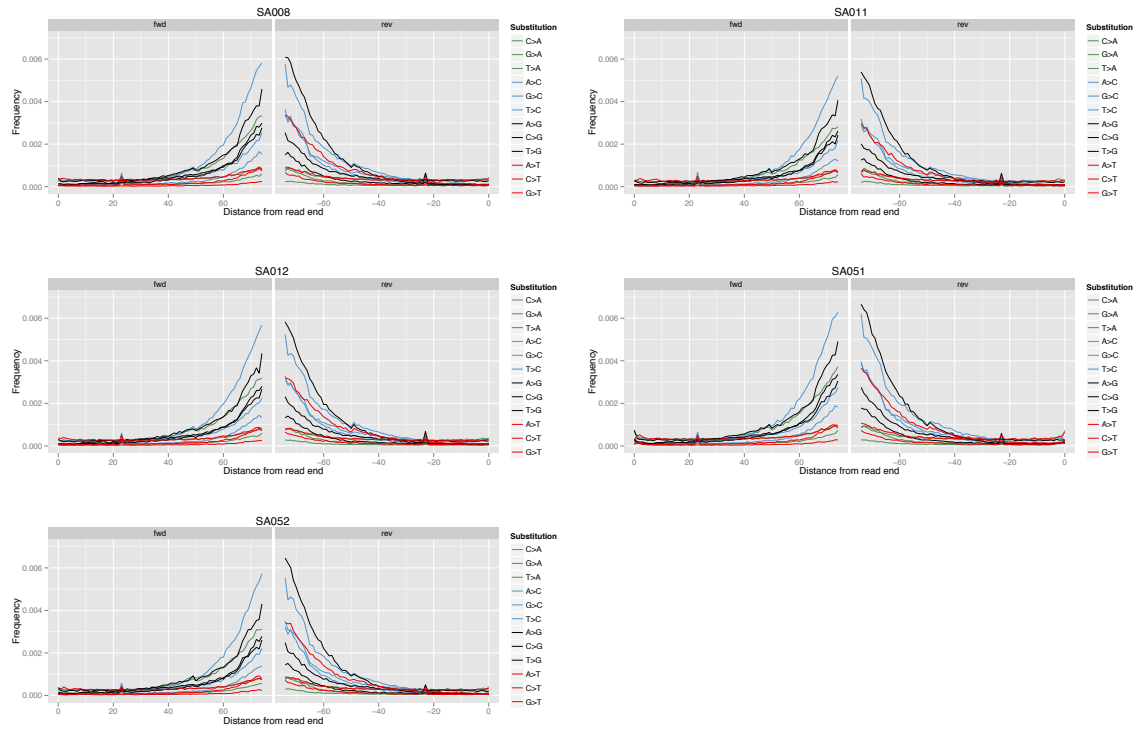


B

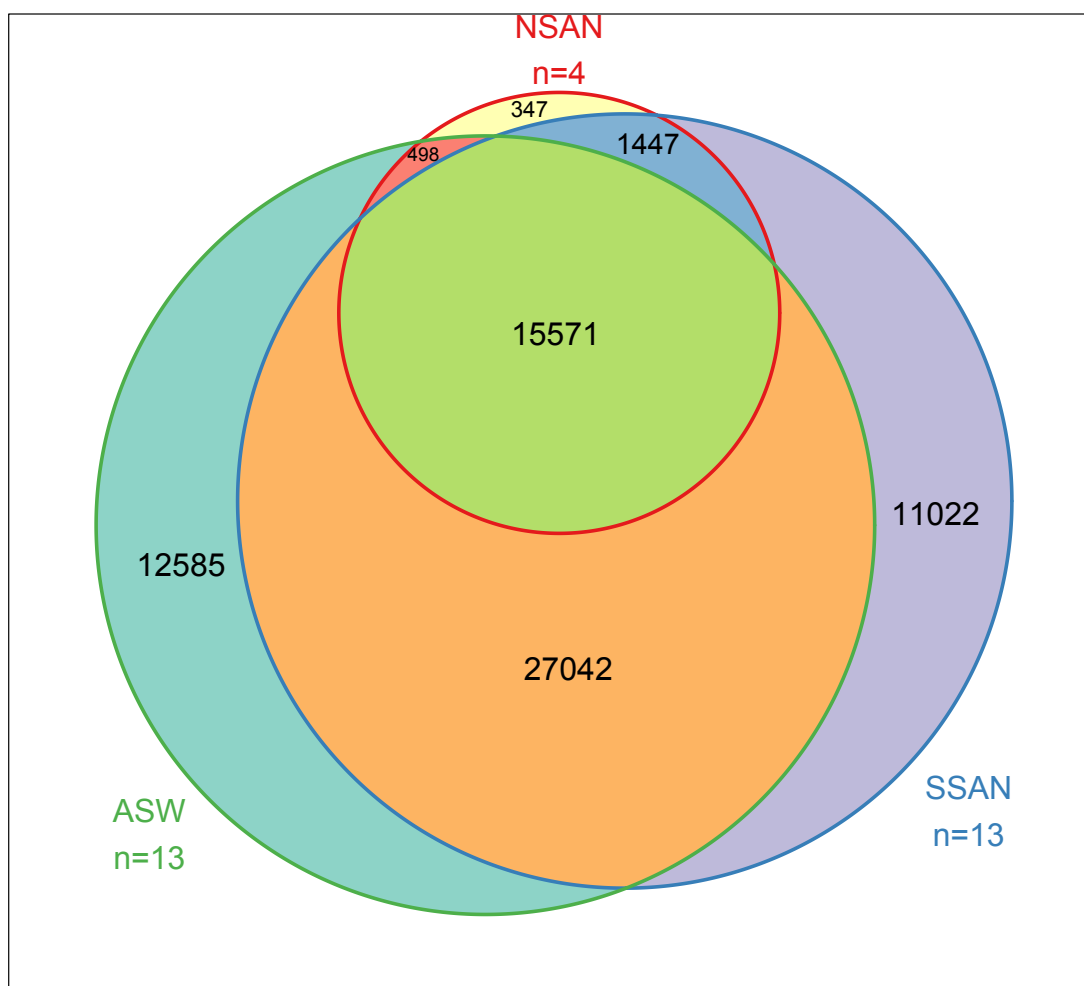


**Figure S3 Mapping quality for all reads.** Mapping quality proportions for all uniquely mapped reads from each individual sample from Pilot 1 (A) and Pilot 2 (B) are shown. Reads included both exome on-target capture and reads mapping off-target to the human reference genome. Mapping quality  $\geq 30$  is shown in red,  $\geq 20$  is orange,  $\geq 10$  is yellow,  $\geq 3$  is green, and  $\leq 3$  is blue.

## SUPPLEMENTARY MATERIAL

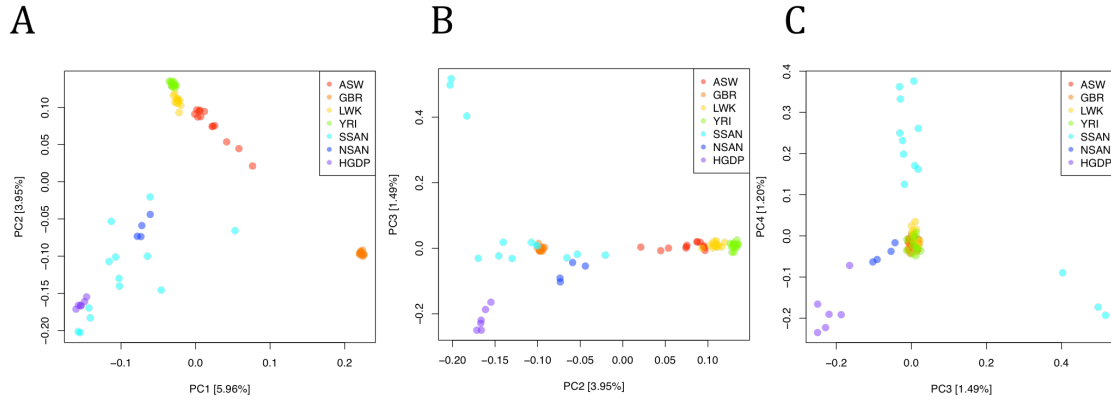


**Figure S4** *Assessment of base substitutions from mapped reads.* Each mapped read was compared to the genome reference sequence to assess patterns consistent with DNA degradation. At each of the 75 positions along a read, we plot the frequency of substitution types, for both the forward (left) and reverse (right) reads from each read-pair. Analysis was limited to 1 million reads from chromosome 1; all raw reads are plotted.



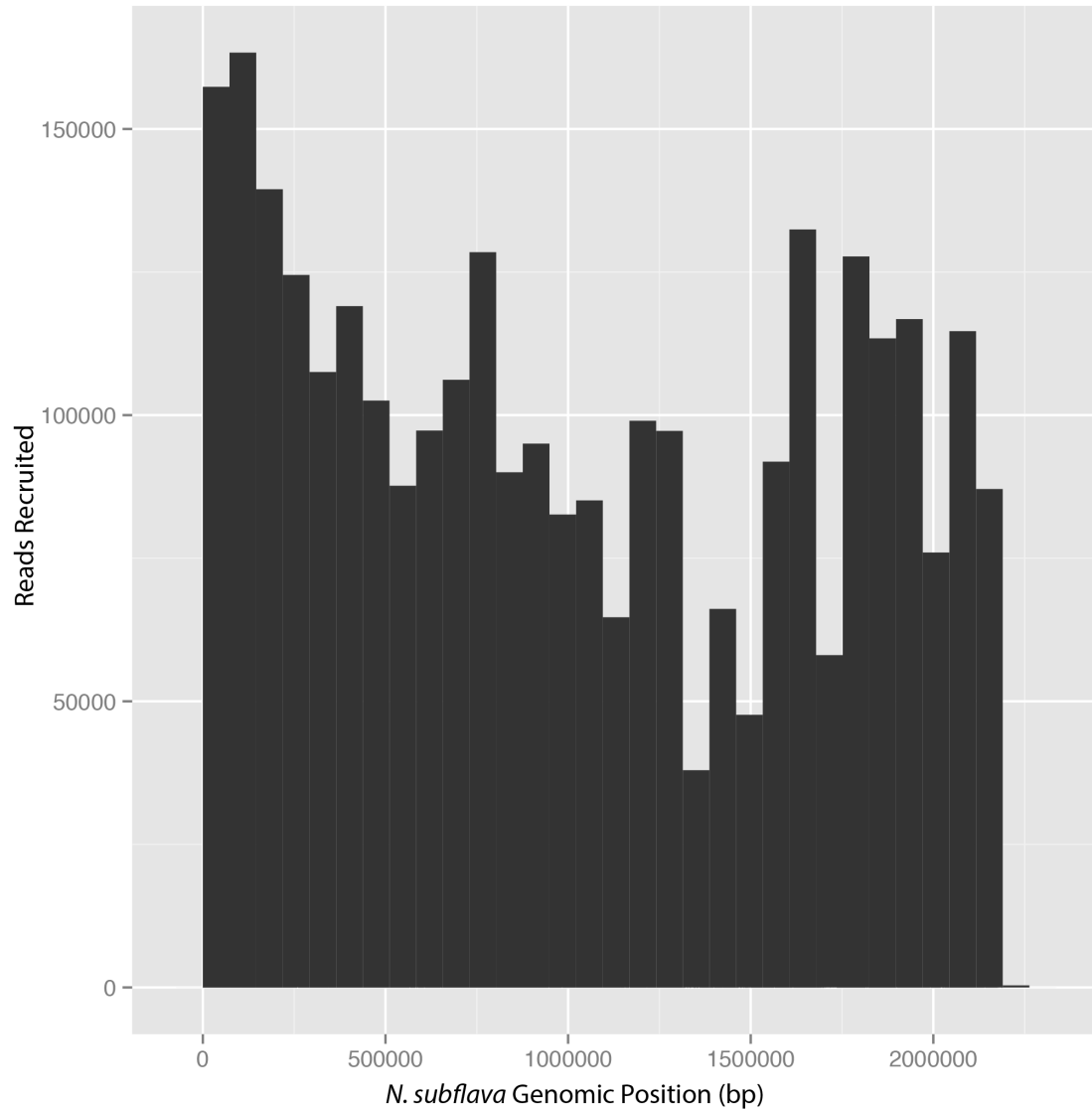
**Figure S5** Venn Diagram comparing  $\neq$  Khomani San with Namibian exome samples. We compared the number of non-reference variants in our South African Khoesans obtained with Agilent capture [SSAN] to the 4 Namibian San exomes from Schuster et al. obtained with NimbleGen capture [NSAN], along with African Americans [ASW] from the 1000 Genomes Project.

## SUPPLEMENTARY MATERIAL



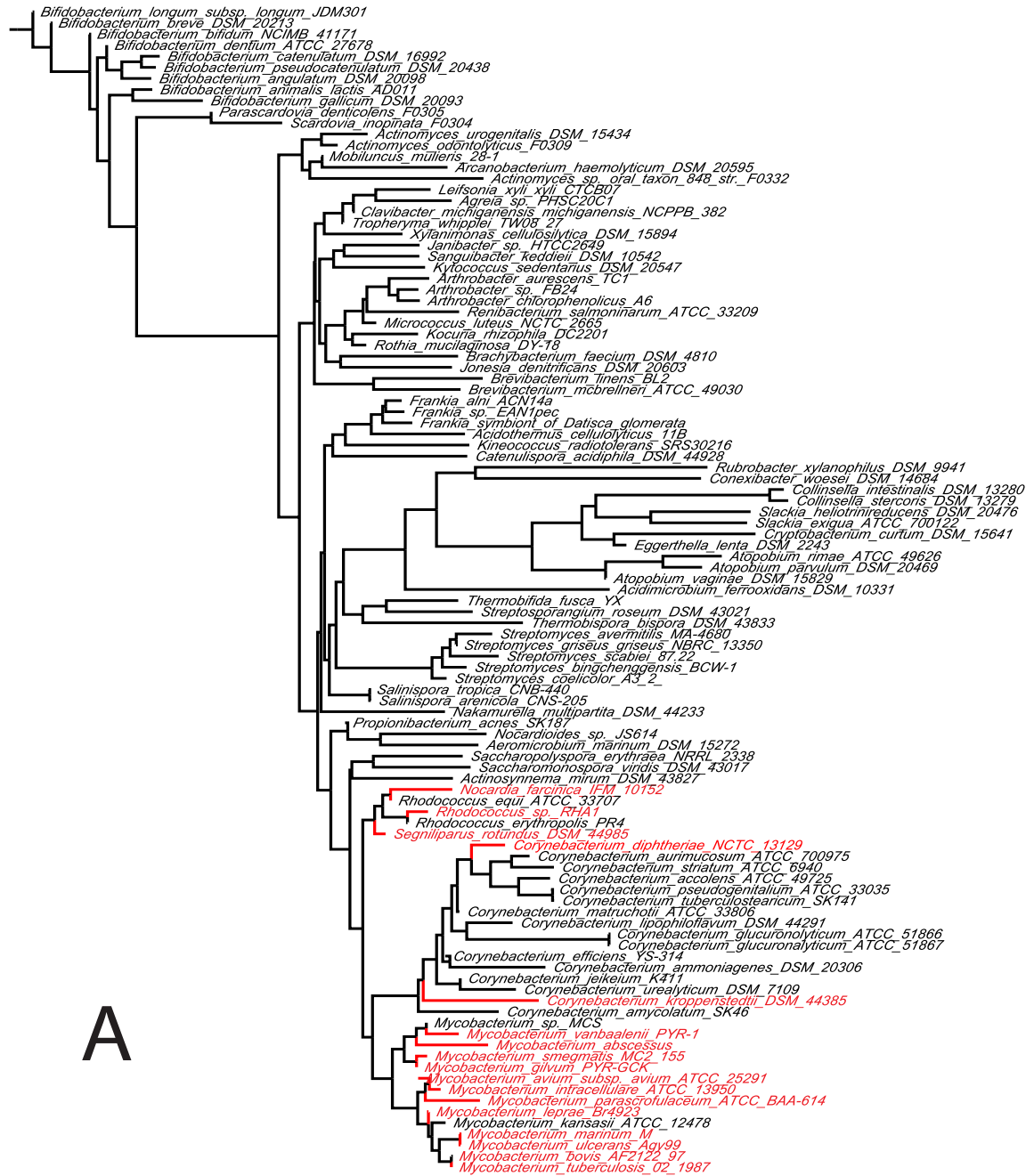
**Figure S6:** *PCA with two relatives included.* We repeated the Principal Component Analysis including two related samples from our  $\neq$ Khomani San population (SA052 and SA054 from Family 2) in order to assess whether SA051 was an outlier. All individuals from Family 2 separate from other  $\neq$ Khomani San along PC3 and PC4, suggesting differences in ancestry likely due to recent population structure in the southern Kalahari.

## SUPPLEMENTARY MATERIAL



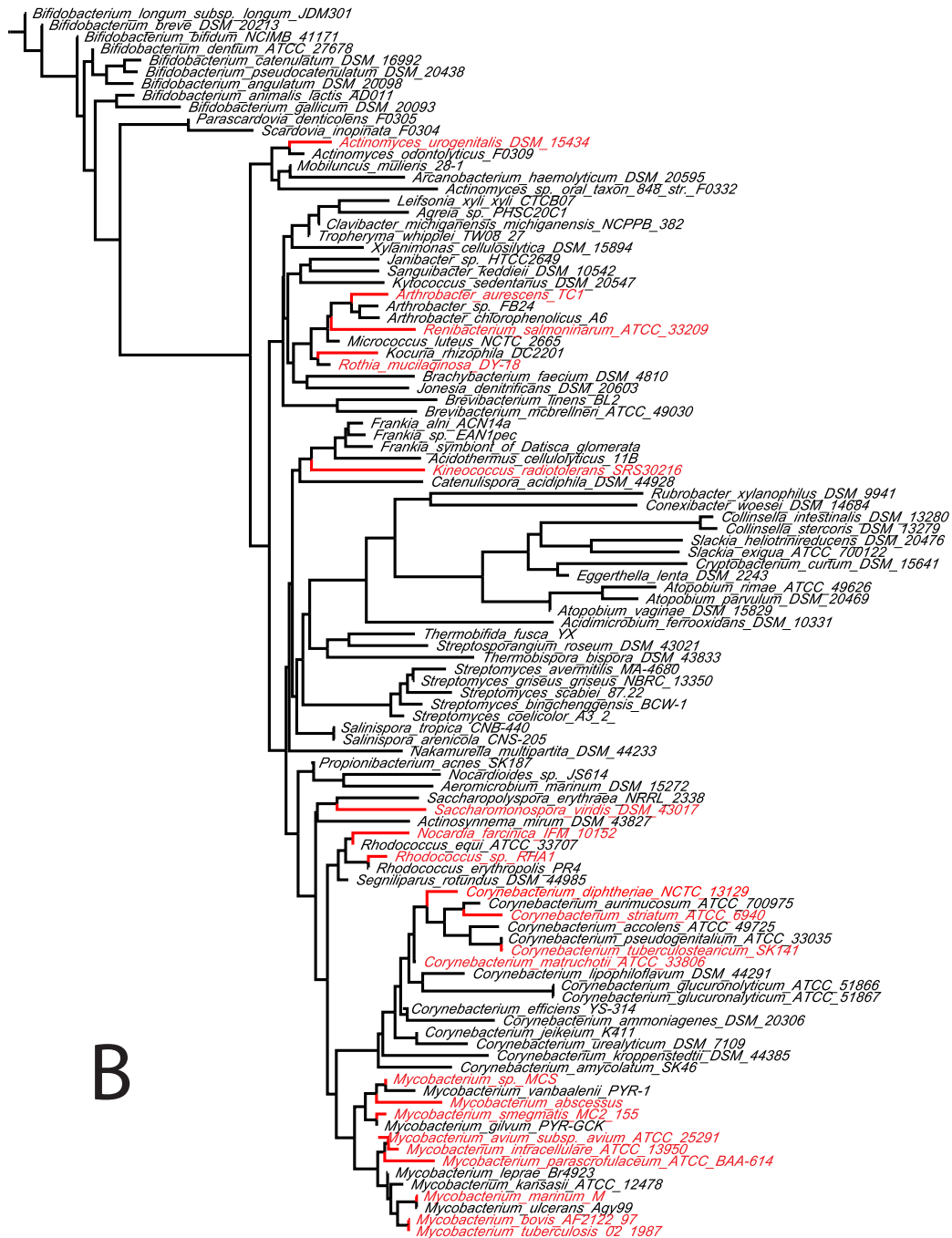
**Figure S7** *Distribution of mapped reads along the *N. subflava* genome.* This histogram illustrates the number of reads that align to various positions of the *N. subflava* genome with 95% identity across at least 75% of the read's length. *N. subflava* genomic contigs were concatenated together and the number of reads that mapped to each base in the concatenated genome were counted. All non-human reads from all KhoeSan samples subject to exome capture were used to generate this histogram.

# SUPPLEMENTARY MATERIAL

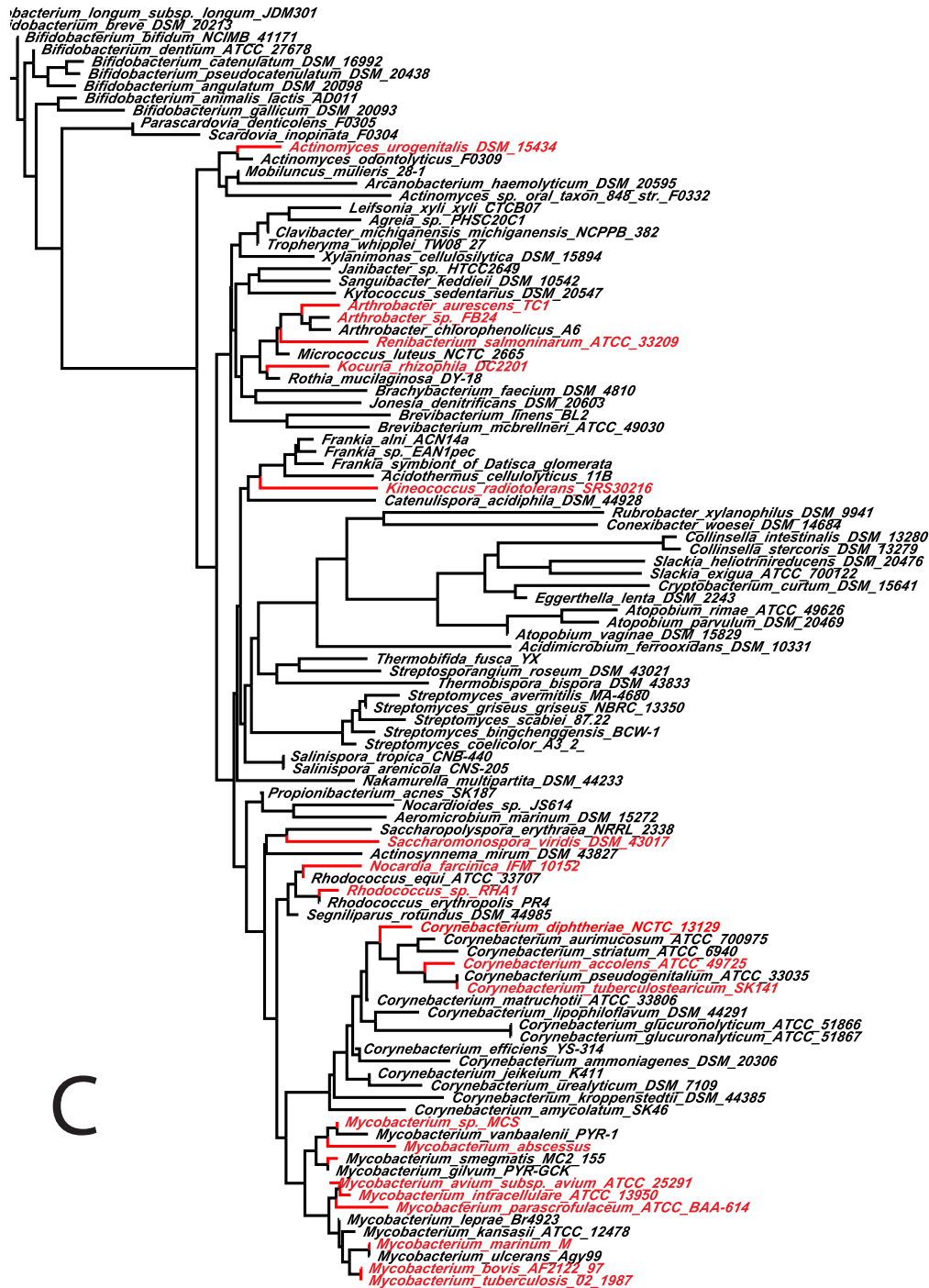




# SUPPLEMENTARY MATERIAL

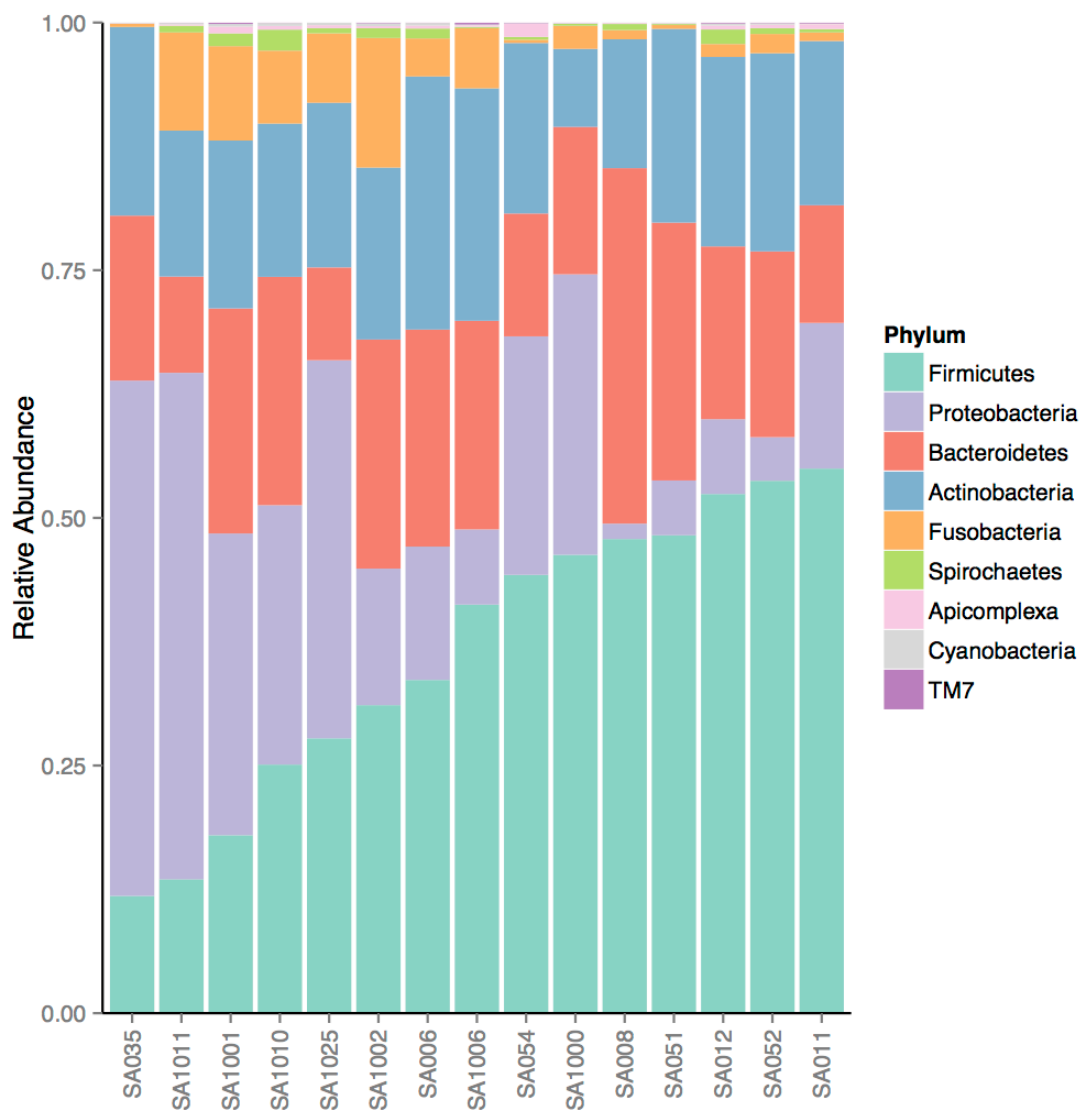


## SUPPLEMENTARY MATERIAL



**Figure S8** The phylogenetic distribution of three non-human exome capture sequences that map with high fidelity to *Mycobacterium tuberculosis*. Each read's distribution is illustrated in the SSU-rRNA (i.e., 16S rRNA) phylogenies above (A, B, and C), which were generated by the Joint Genomes Institutes via the IMG database using PHYLIP. Lineages colored in red represent species whose genome recruit the read with equal fidelity as the *M. tuberculosis* genome. Lineages colored in black represent reference genomes that did not recruit the read. Only Actinobacteria are included in these phylogenies.

## SUPPLEMENTARY MATERIAL



**Figure S9** The phylum-level structure of the oral microbiome structure varies among the Khoesan. Each of the above stacked bar plots illustrates the relative abundance (y-axis) of the most abundant oral phyla for each of the 15 Khoesan individuals (x-axis). Relative abundance was measured as the fraction of high-quality reads that were recruited to a microbial genome of a particular taxonomic rank using conservative recruitment settings (Methods).